Normal Gamma model

Kevin P. Murphy
murphyk@cs.ubc.ca

Last updated October 1, 2007

0.1 Normal-Gamma model

In this section, we consider the case where the mean and precision are both unknown. We just state the results without proofs. Derivations may be found in [Mur07]. First we introduce two useful distributions.

0.1.1 Gamma distribution

The gamma distribution is a flexible distribution for positive real valued rv's, $x > 0$. It is defined in terms of two parameters. There are two common parameterizations. This is the one used by Bishop [Bis06] (and many other authors):

$$Ga(x|\text{shape} = a, \text{rate} = b) \quad = \quad \frac{b^a}{\Gamma(a)} x^{a-1} e^{-xb}, \quad x, a, b > 0 \tag{1}$$

The second parameterization (and the one used by Matlab's `gampdf`) is

$$Ga(x|\text{shape} = \alpha, \text{scale} = \beta) \quad = \quad \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} \tag{2}$$

Note that the shape parameter controls the shape; the scale parameter merely defines the measurement scale (the horizontal axis). The rate parameter is just the inverse of the scale. See Figure 1 for some examples. This distribution has the following properties (using the rate parameterization):

$$\text{mean} \quad = \quad \frac{a}{b} \tag{3}$$

$$\text{mode} \quad = \quad \frac{a-1}{b} \text{ for } a \geq 1 \tag{4}$$

$$\text{var} \quad = \quad \frac{a}{b^2} \tag{5}$$

0.1.2 Student t distribution

The generalized t-distribution is given as

$$t_\nu(x|\mu, \sigma^2) \quad = \quad c \left[1 + \frac{1}{\nu} (\frac{x-\mu}{\sigma})^2 \right]^{-(\frac{\nu+1}{2})} \tag{6}$$

$$c \quad = \quad \frac{\Gamma(\nu/2 + 1/2)}{\Gamma(\nu/2)} \frac{1}{\sqrt{\nu\pi}\sigma} \tag{7}$$

where c is the normalization consant. μ is the mean, $\nu > 0$ is the **degrees of freedom**, and $\sigma^2 > 0$ is the scale. (Note that the ν parameter is written as a subscript.)

The distribution has the following properties:

$$\text{mean} \quad = \quad \mu, \ \nu > 1 \tag{8}$$

$$\text{mode} \quad = \quad \mu \tag{9}$$

$$\text{var} \quad = \quad \frac{\nu\sigma^2}{(\nu-2)}, \quad \nu > 2 \tag{10}$$

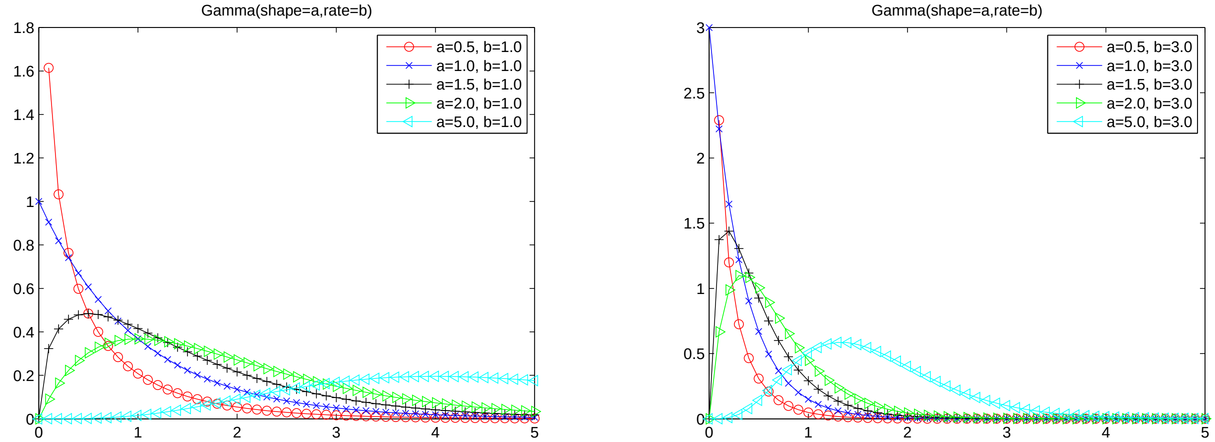

Figure 1: Some $Ga(a, b)$ distributions. If $a < 1$, the peak is at 0. As we increase b, we squeeze everything leftwards and upwards. Figures generated by `gammaDistPlot2`.

Note: if $x \sim t_\nu(\mu, \sigma^2)$, then

$$\frac{x - \mu}{\sigma} \sim t_\nu \tag{11}$$

which corresponds to a standard t-distribution with $\mu = 0, \sigma^2 = 1$ (Matlab's `tpdf`):

$$t_\nu(x) = \frac{\Gamma((\nu + 1)/2)}{\sqrt{\nu\pi}\Gamma(\nu/2)}(1 + x^2/\nu)^{-(\nu+1)/2} \tag{12}$$

In Figure 2, we plot the density for different parameter values. T-distributions are like Gaussian distributions with **heavy tails**. Hence they are more robust to outliers (see Figure 3). As $\nu \to \infty$, the T approaches a Gaussian.

If $\nu = 1$, this is called a **Cauchy distribution**. This is an interesting distribution since if $X \sim Cauchy$, then $E[X]$ does not exist, since the corresponding integral diverges. Essentially this is because the tails are so heavy that samples from the distribution can get very far from the center μ.

It can be shown that the t-distribution is like an infinite sum of Gaussians, where each Gaussian has a different variance [Arc05, p111]:

$$t_\nu(x|\mu, \lambda^{-1}) = \int_0^\infty \mathcal{N}(x|\mu, (u\lambda)^{-1}) Ga(u|\text{shape}=\frac{\nu}{2}, \text{rate}=\frac{\nu}{2}) du \tag{13}$$

(See exercise 2.46 of [Bis06].)

0.2 Likelihood

The likelihood can be written in this form

$$p(D|\mu, \lambda) = \frac{1}{(2\pi)^{n/2}} \lambda^{n/2} \exp\left(-\frac{\lambda}{2}\sum_{i=1}^n (x_i - \mu)^2\right) \tag{14}$$

$$= \frac{1}{(2\pi)^{n/2}} \lambda^{n/2} \exp\left(-\frac{\lambda}{2}\left[n(\mu - \overline{x})^2 + \sum_{i=1}^n (x_i - \overline{x})^2\right]\right) \tag{15}$$

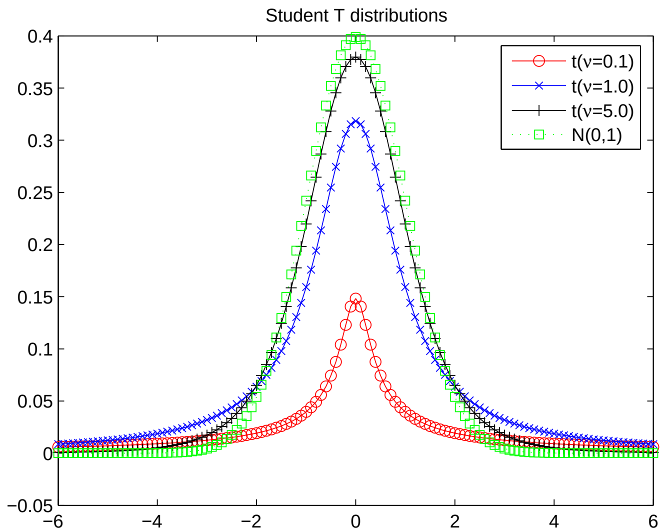

Figure 2: Student t-distributions $T_\nu(\mu, \sigma^2)$ for $\mu = 0$. The effect of σ is just to scale the horizontal axis. As $\nu \rightarrow \infty$, the distribution approaches a Gaussian. See `studentTplot`.

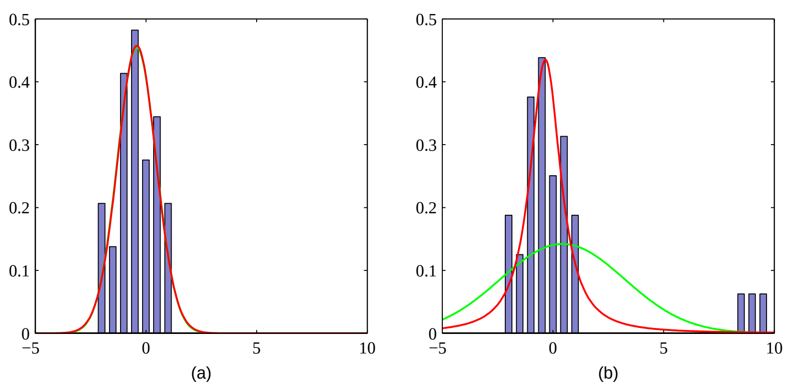

Figure 3: Fitting a Gaussian and a Student distribution to some data (left) and to some data with outliers (right). The Student distribution (red) is much less affected by outliers than the Gaussian (green). Source: [Bis06] Figure 2.16.

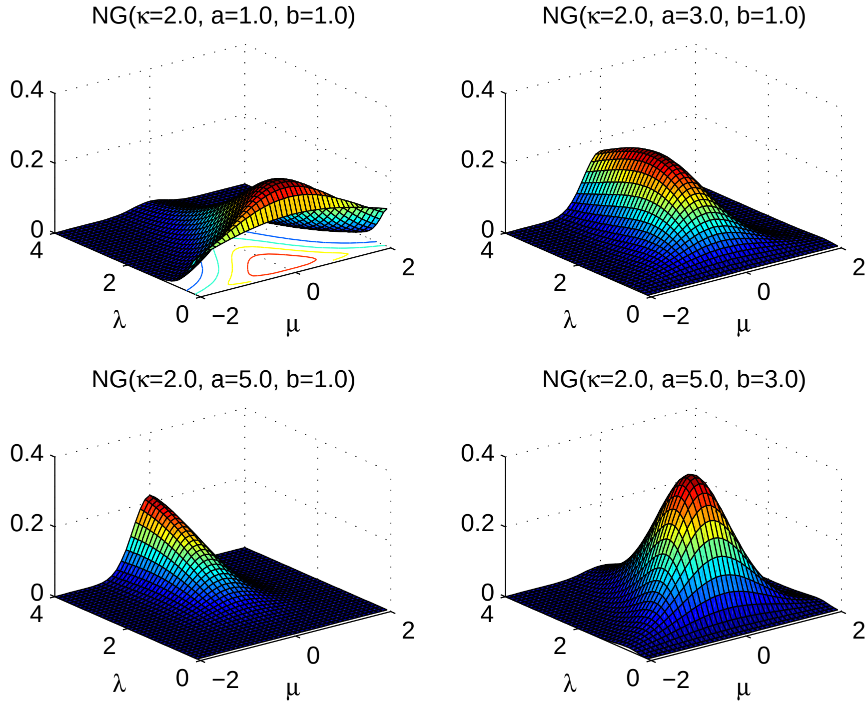

Figure 4: Some Normal-Gamma distributions. Produced by `NGplot2`.

0.3 Prior

The conjugate prior is the **normal-Gamma**:

$$NG(\mu, \lambda|\mu_0, \kappa_0, \alpha_0, \beta_0) \stackrel{\text{def}}{=} \mathcal{N}(\mu|\mu_0, (\kappa_0\lambda)^{-1})Ga(\lambda|\alpha_0, \text{rate} = \beta_0) \quad (16)$$

$$= \frac{1}{Z_{NG}(\mu_0, \kappa_0, \alpha_0, \beta_0)}\lambda^{\frac{1}{2}} \exp(-\frac{\kappa_0\lambda}{2}(\mu - \mu_0)^2)\lambda^{\alpha_0-1}e^{-\lambda\beta_0} \quad (17)$$

$$= \frac{1}{Z_{NG}}\lambda^{\alpha_0-\frac{1}{2}} \exp\left(-\frac{\lambda}{2}\left[\kappa_0(\mu - \mu_0)^2 + 2\beta_0\right]\right) \quad (18)$$

$$Z_{NG}(\mu_0, \kappa_0, \alpha_0, \beta_0) = \frac{\Gamma(\alpha_0)}{\beta_0^{\alpha_0}}\left(\frac{2\pi}{\kappa_0}\right)^{\frac{1}{2}} \quad (19)$$

See Figure 4 for some plots.

0.4 Posterior

The posterior is

$$p(\mu, \lambda|D) = NG(\mu, \lambda|\mu_n, \kappa_n, \alpha_n, \beta_n) \quad (20)$$

$$\mu_n = \frac{\kappa_0\mu_0 + n\overline{x}}{\kappa_0 + n} \quad (21)$$

$$\kappa_n = \kappa_0 + n \quad (22)$$

$$\alpha_n = \alpha_0 + n/2 \quad (23)$$

$$\beta_n = \beta_0 + \tfrac{1}{2}\sum_{i=1}^{n}(x_i - \overline{x})^2 + \frac{\kappa_0 n(\overline{x} - \mu_0)^2}{2(\kappa_0 + n)} \quad (24)$$

We see that the posterior sum of squares, β_n, combines the prior sum of squares, β_0, the sample sum of squares, $\sum_i (x_i - \overline{x})^2$, and a term due to the discrepancy between the prior mean and sample mean. As can be seen from Figure 4, the range of probable values for μ and σ^2 can be quite large even after for moderate n. Keep this picture in mind whenever someones claims to have "fit a Gaussian" to their data.

The posterior marginals are

$$\begin{aligned} p(\lambda|D) &= Ga(\lambda|\alpha_n, \beta_n) && (25) \\ p(\mu|D) &= T_{2\alpha_n}(\mu|\mu_n, \frac{\beta_n}{\alpha_n \kappa_n}) && (26) \end{aligned}$$

0.5 Marginal likelihood

$$\begin{aligned} p(D) &= \frac{Z_n}{Z_0}(2\pi)^{-n/2} && (27) \\ &= \frac{\Gamma(\alpha_n)}{\Gamma(\alpha_0)} \frac{\beta_0^{\alpha_0}}{\beta_n^{\alpha_n}} (\frac{\kappa_0}{\kappa_n})^{\frac{1}{2}} (2\pi)^{-n/2} && (28) \end{aligned}$$

0.6 Posterior predictive

$$p(x|D) = t_{2\alpha_n}(x|\mu_n, \frac{\beta_n(\kappa_n + 1)}{\alpha_n \kappa_n}) \quad (29)$$

References

[Arc05] C. Archamebau. *Probabilistic models in noisy environments*. PhD thesis, U. Catholique de Louvain, Machine learning group, 2005.

[Bis06] C. Bishop. *Pattern recognition and machine learning*. Springer, 2006.

[Mur07] K. Murphy. Conjugate bayesian analysis of the gaussian distribution. Technical report, UBC, 2007.